氏　　名	大前　勝弘
学位(専攻分野)	博士(統計科学)
学 位 記 番 号	総研大甲第 1994 号
学位授与の日付	平成 30 年 3 月 23 日
学位授与の要件	複合科学研究科　統計科学専攻 学位規則第6条第1項該当
学 位 論 文 題 目	Statistical Learning by Quasi-linear Predictor
論 文 審 査 委 員	主　査　准教授　間野　修平 教授　江口　真透 准教授　野間　久史 教授　松浦　正明　帝京大学　大学院 公衆衛生学研究科

(別紙様式 2)
(Separate Form 2)

論文の要旨
Summary (Abstract) of doctoral thesis contents

In a biomedical study, a linear predictor is widely used as the regression and prediction function because of the learnability and understandability. Statistical models are usually formulated by the linear predictor. Fisher's linear discriminant analysis is a representative example of such formulations for a binary classification. It has the Bayes risk consistency in the classification problem of two normal samples with equal variances. The linear form is thus useful in fitting to data with homogenous structure. However, especially in the field of biomedical science, it has been revealed that there are several diseases once considered to be homotypic but later elucidated as heterotypic. Nevertheless, only a linear model has been applied and the estimated model is used for following discussion in most biomedical studies. This fact may yield misleading for therapy by failing to reveal the potentially complex mixture of substantial benefits and harms. Such heterogeneous diseases do not allow only a single set of the biomarkers to be predictive for all patients. Moreover, the heterogeneity may obstruct us to detect the predictive biomarkers and to learn the predictive model. We therefore need to construct a predictor in consideration of the heterogeneous structure.

Taking such a background into account, the author derived the quasi-linear predictor defined as log-sum-exp form. It is a special case of the generalized average known as the Kolmogorov-Nagumo average. The quasi-linear predictor is made up of the combination of some liner predictors with the different intercepts and coefficients. The shape of the quasi-liner predictor is determined not only by the parameters of these linear predictors but also by the tuning parameter for adjusting the overall nonlinearity. The quasi-linear predictor converges to the minimum and maximum of the linear predictors, and reduces to the linear predictor, in the limiting sense of the tuning parameter. For analytical purpose, the tuning parameter is determined by the Bayes information criteria with the learning dataset. It results in that the suitable non-linearity of the regression or prediction function is estimated by the data.

In this thesis, the author extended two ordinary models, the linear logistic model and Cox's proportional hazard model, to the quasi-linear logistic model and the quasi-linear relative risk model, respectively. The optimality of the quasi-linear logistic model is assured when we consider the binary classification problem of mixture normal sample and normal sample with equal variances for each component. To get more parsimonious expression, the author derived the restricted quasi-linear logistic model, which is defined as the logistic model with the quasi-linear predictor, where each predictor is composed by known disjoint clusters of covariates. The restricted model gives easier interpretation of the estimated model compared to the regular quasi-linear model, because each covariate is incorporated in only one of the

linear predictors. Moreover, for the case such clusters are unknown, the author derived the quasi-linear logistic model with the cross-L1 regularization. In this regularization method, a penalty for product of absolute values of coefficients for same covariate in different clusters is added to the likelihood function. Sufficiently strong cross-L1 penalty therefore results in the restricted quasi-linear model and the resultant disjoint sets of covariates are automatically given through the model estimation procedure. The second model, the quasi-linear relative risk model, is regarded as an extension of a mixture hazard model. In fact, the mixture hazard model with same baseline hazard function corresponds to the quasi-linear relative risk model with a specific tuning parameter. As is the case with the quasi-linear logistic model, the quasi-linear relative risk model with the cross-L1 regularization is derived. The extensions of these two models were performed simply by replacing the linear predictor to the quasi-linear predictor. Other extensions of linear models are easily combined and implemented. The author derived the L1 and L2 penalized versions for both of the quasi-linear logistic model and the quasi-linear relative risk models.

In the simulation study for the binary classification problem, the author checked the consistency of the parameter estimation and compared the predictive performance between the linear logistic model and the restricted quasi-linear logistic model. In the application studies for the binary classification problem, the author compared the performance among the restricted quasi-linear logistic model, linear logistic model and ordinary classification methods including decision tree, random forest, support vector machine, naive Bayes, group lasso, neural network, L1 and L2 penalized linear logistic models. These simulation and application studies showed that the restricted quasi-linear logistic model has better performance in some simulated examples and real datasets than the ordinary methods. For the regression problem on the survival time data, the author checked the true model selection probability by the Bayes information criteria and the parameter consistency for some situations in the simulation studies and compared the quasi-linear relative risk model with cross L1 penalty and Cox's proportional hazard model in the application studies. The simulation studies showed that the parameter estimation empirically has the consistency and the selection of the tuning parameter by the Bayes information criteria works very well. The application studies for the regression problem on the survival time data showed that the quasi-linear relative risk model has better performance in some real datasets compared to the Cox's proportional hazard model.

Finally, the author discussed the role of the quasi-linear predictor in traditional clustering methods, and the relationship among the quasi-linear model, mixture of experts model, and neural network model for more discussions and future works.

(別紙様式 3)
(Separate Form 3)

博士論文審査結果の要旨
Summary of the results of the doctoral thesis screening

大前勝弘氏の論文発表会と審査委員会を、平成 30 年 1 月 19 日午後 4 時から約 1 時間 30 分に渡り、審査委員全員の出席のもとに開催した。出願者による概要説明と質疑応答による公開の論文発表会の後、審査員のみによる審査と口述による試験を行った。

提出された博士出願論文は英語で全 6 章 106 ページから成り、複数の線形モデルを統合する準線形予測モデルに関して論じている。準線形予測モデルは共変量の異質性のモデル化のために一般化平均により線形予測子を非線形に結合した準線形予測子に基づく。モデルの過剰な複雑化を防ぐために共変量にクラスター分析を援用した制限モデルを提案し、モデルの識別性を担保するために各線形予測子への共変量の寄与が互いに素に近づくようにペナルティを導入した Lasso タイプの正則化を導入する工夫を行っている。

第 1 章では、統計的予測問題の文献レビューを行い、従来の線形ロジスティックモデルにおいては共変量に潜在する異質性の問題にうまく対処できていないことを指摘し、それが本論文での提案手法の動機付けとなっていることを述べている。

第 2 章では、本論文で提案する新たな手法について論じている。まず、典型的な対数和指数と呼ばれる平均を使って準線形予測子を定義し、その予測子としての解析的性質をまとめている。さらに幾つかの一般化平均を使った準線形予測子を考察している。

第 3 章では、主要結果を述べる準備のために、標準的なモデルと統計分析、特に、一般化線形モデル、相対リスクモデルの尤度関数、尤度方程式について論じている。

第 4 章では、準線形ロジスティック回帰モデルのベイズリスク一致性について考察している。コントロール群に正規分布、ケース群に正規混合モデルを仮定し、さらに等分散性を仮定すると、準線形予測子が対数尤度比統計量に等しいことから、準線形予測子がケース群の異質性によって生じた混合分布を素直にモデル化していることを示している。幾つかの数値実験と実データ解析から提案手法の有用性を示している。

第 5 章では、準線形相対リスクモデルについて考察している。比例ハザード性を仮定すると準線形予測子の標準化が必要なことを指摘し、その変換の下で偏尤度解析についての定式化を行っている。さらに、共変量選択とモデルの識別性を考慮した Lasso タイプの正則化を提案している。数値実験と実データ解析から標準的なコックス回帰モデルとの比較によって提案手法の有用性を結論している。

第 6 章では、準線形モデルの関連する研究のサーベイを行い、今後の研究の発展方向について考察している。とくにクラスター分析について、最大エントロピー・クラスタリングと k -means を含む一般化が可能であることを指摘している。

以上に述べたように、提出された博士出願論文は、予測問題の重要な統計学的課題を深く考察し、従来の手法を超えた興味深い提案を行い、それを明快に説明している。実際、本研究で提案されたデータの異質性を柔軟に扱う手法は、この課題に大きく貢献すると思われる。得られた主要結果については既に査読付き学術雑誌に掲載されている。これらの理由により本博士出願論文は優れた論文であると考えられるため、審査委員会は、本博士出願論文は学位授与の水準に達していると判断し、本博士論文審査の合格を推薦する結論に達した。